

A short note on the tail bound of Wishart distribution

Shenghuo Zhu
zsh@nec-labs.com

December 27, 2012

Abstract

We study the tail bound of the empirical covariance of multivariate normal distribution. Following the work of (Gittens & Tropp, 2011), we provide a tail bound with a small constant.

1 Main result

Let $\{\xi_k : k = 1 \cdots n\}$ follow multivariate normal distribution $\mathcal{N}_d(0, C)$. The scatter matrix $S = \sum_{k=1}^n \xi_k \xi_k^\top$ follows Wishart distribution, $\mathcal{W}_d(n, C)$. The estimate of C is $\frac{1}{n}S$. The tail bound of S has a wide range of applications, such as, the sample estimation of random projection. We follow the work of (Gittens & Tropp, 2011) to find the tail bound with smaller constants.

Notation: Let denote the ℓ -th largest eigenvalue of matrix X by $\lambda_\ell(X)$, the trace of X by $\text{tr}(X)$, and the spectral norm of X by $\|X\|$.

Theorem 1. *If S follows a Wishart distribution $\mathcal{W}_d(n, C)$, then for $\theta \geq 0$,*

$$\Pr \left\{ \lambda_1\left(\frac{1}{n}S - C\right) \geq \left(\sqrt{\frac{2\theta(r+1)}{n}} + \frac{2\theta r}{n} \right) \lambda_1(C) \right\} \leq d \exp\{-\theta\}, \quad (1)$$

$$\Pr \left\{ \lambda_1\left(C - \frac{1}{n}S\right) \geq \left(\sqrt{\frac{2\theta(r+1)}{n}} + \frac{2\theta r}{n} \right) \lambda_1(C) \right\} \leq d \exp\{-\theta\}, \quad (2)$$

$$\Pr \left\{ \left\| \frac{1}{n}S - C \right\| \geq \left(\sqrt{\frac{2\theta(r+1)}{n}} + \frac{2\theta r}{n} \right) \|C\| \right\} \leq 2d \exp\{-\theta\}, \quad (3)$$

$$\Pr \left\{ \left| \lambda_\ell\left(\frac{1}{n}S\right) - \lambda_\ell(C) \right| \geq \left(\sqrt{\frac{2\theta\kappa_\ell^2(r+1)}{n}} + \frac{2\theta\kappa_\ell r}{n} \right) \lambda_\ell(C), \forall \ell \in \{1 \cdots d\} \right\} \leq 2d \exp\{-\theta\}, \quad (4)$$

where $r = \text{tr}(C)/\|C\|$, and condition numbers $\kappa_\ell = \lambda_1(C)/\lambda_\ell(C)$.

Remark 1. When $d = 1$ and $C = 1$, then $r = 1$, and it is exactly the upper bound of chi-square distribution provided in in (Laurent & Massart, 2000).

Remark 2. Applying the modification in this note to Theorem 7.1 of (Gittens & Tropp, 2011), we have

$$\Pr \left\{ \lambda_\ell\left(\frac{1}{n}S\right) \geq \left(1 + \sqrt{\frac{2\theta(\kappa_\ell r_\ell + 2)}{n}} + \frac{2\theta\kappa_\ell r_\ell}{n} \right) \lambda_\ell(C) \right\} \leq (d - \ell + 1) \exp\{-\theta\}, \text{ for } \ell = 1 \cdots d, \quad (5)$$

$$\Pr \left\{ \lambda_\ell\left(\frac{1}{n}S\right) \leq \left(1 - \sqrt{\frac{2\theta\kappa_\ell^2(r_1 - r_{\ell+1} + 2)}{n}} \right) \lambda_\ell(C) \right\} \leq \ell \exp\{-\theta\}, \text{ for } \ell = 1 \cdots d, \quad (6)$$

where $r_\ell = \sum_{i=\ell}^d \lambda_i(C)/\lambda_1(C)$. As r_ℓ is smaller than r , it is tighter individually. Eq. (5) and (6) are individual eigenvalue bounds, but Eq. (4) is the collective eigenvalue bound. When $\ell = 1$, $\kappa_1 = 1$ and $r_\ell = r$, then the upper bound of the top eigenvalue of Eq. (5) is slightly looser than that of Eq. (1).

2 Proof

We use part of the proof of Lemma 8 in (Birgé & Massart, 1998).

Lemma 2. *Let $B > 0$ and $\sigma > 0$. If the log-moment generating function satisfies*

$$\log \mathbb{E} \exp\{uZ\} \leq \frac{\sigma^2 u^2}{2(1-uB)} \quad \text{for all } 0 \leq u < 1/B,$$

then

$$\Pr\{Z \geq \epsilon\} \leq \exp\left\{-\frac{\epsilon^2}{2\sigma^2 + 2\epsilon B}\right\} \quad \text{for all } \epsilon \geq 0, \quad (7)$$

and

$$\Pr\{Z \geq \sqrt{2\theta\sigma^2} + \theta B\} \leq \exp\{-\theta\} \quad \text{for all } \theta \geq 0. \quad (8)$$

Proof. It follows Markov's inequality that

$$\Pr\{Z \geq \epsilon\} \leq \inf_u \mathbb{E} \exp\{-u\epsilon + uZ\} = \exp\{-h(\epsilon)\},$$

where $h(\epsilon) := \sup_u u\epsilon - \frac{\sigma^2 u^2}{2(1-uB)}$. Also, the supremum is achieved for

$$\epsilon = \frac{\sigma^2 u}{1-uB} + \frac{\sigma^2 u^2 B}{2(1-uB)^2} = \frac{\sigma^2 u}{2(1-uB)} + \frac{\sigma^2 u}{2(1-uB)^2},$$

i.e. $u = B^{-1}[1 - \sigma(2\epsilon B + \sigma^2)^{-1/2}] < 1/B$. Then we prove Eq. (7), as

$$h(\epsilon) = \frac{\epsilon^2}{\epsilon B + \sigma^2 + \sigma^2(1 + 2\epsilon B/\sigma^2)^{1/2}} \geq \frac{\epsilon^2}{2\epsilon B + 2\sigma^2}.$$

Let

$$\theta := \frac{\sigma^2 u^2}{2(1-uB)^2} = h(\epsilon).$$

Then we prove Eq. (8), as

$$\sqrt{2\theta\sigma^2} + \theta B = \frac{\sigma^2 u}{(1-uB)} + \frac{\sigma^2 u^2 B}{2(1-uB)^2} = \epsilon.$$

□

The following Theorem is Theorem 6.2 in (Tropp, 2010), except for using Lemma 2 to achieve a different formula.

Theorem 3. *If a finite sequence $\{X_k : k = 1 \dots n\}$ of independent, random, self-adjoint matrices with dimension d , all of which satisfy the Bernstein's moment condition, i.e.*

$$\mathbb{E} X_k^p \preceq \frac{p!}{2} B^{p-2} \Sigma_2, \quad \text{for } p \geq 2,$$

where B is a positive constant and Σ_2 is a positive semi-definite matrix, then,

$$\begin{aligned} \log \mathbb{E} \exp(uX_k) &\preceq u\mathbb{E} X_k + \frac{u^2}{2(1-uB)} \Sigma_2 \quad \text{for all } 0 \leq u < 1/B, \\ \Pr\{\lambda_1(\sum_k X_k) \geq \lambda_1(\sum_k \mathbb{E} X_k) + \sqrt{2n\theta\lambda_1(\Sigma_2)} + \theta B\} &\leq d \exp\{-\theta\}. \end{aligned}$$

Additionally, if X_k are positive semi-definite matrices,

$$\begin{aligned} \log \mathbb{E} \exp(-uX_k) &\preceq -u\mathbb{E} X_k + \frac{u^2}{2} \Sigma_2 \quad \text{for all } u \geq 0, \\ \Pr\{\lambda_d(\sum_k X_k) \leq \lambda_d(\sum_k \mathbb{E} X_k) - \sqrt{2\theta n\lambda_1(\Sigma_2)}\} &\leq d \exp\{-\theta\}. \end{aligned}$$

Proof.

$$\begin{aligned}
\log \mathbb{E} \exp(uX_k) &= \log(I + u\mathbb{E}X_k + \sum_{p=2}^{\infty} \frac{u^p}{p!} \mathbb{E}X_k^p) \\
&\preceq u\mathbb{E}X_k + \sum_{p=2}^{\infty} \frac{u^2(uB)^{p-2}}{2} \Sigma_2 \\
&= u\mathbb{E}X_k + \frac{u^2}{2(1-uB)} \Sigma_2.
\end{aligned}$$

It follows Theorem 3.6 in (Tropp, 2010) and Lemma 2 that

$$\begin{aligned}
\Pr\{\lambda_1(\sum_k X_k) \geq \lambda_1(\sum_k \mathbb{E}X_k) + \epsilon\} &\leq \inf_{u \geq 0} \left\{ \exp(-u\lambda_1(\sum_k \mathbb{E}X_k) - u\epsilon) \text{tr} \exp(\sum_k \log \mathbb{E} \exp(uX_k)) \right\} \\
&\leq \inf_{u \geq 0} \left\{ d \exp(-u\epsilon + \frac{nu^2}{2(1-uB)} \lambda_1(\Sigma_2)) \right\} \\
&\leq d \exp(-\theta),
\end{aligned}$$

where $\epsilon = \sqrt{2n\theta\lambda_1(\Sigma_2)} + \theta B$.

$$\begin{aligned}
\log \mathbb{E} \exp(-uX_k) &\preceq \log(I - u\mathbb{E}X_k + \frac{u^2}{2} \mathbb{E}X_k^2) \\
&\preceq -u\mathbb{E}X_k + \frac{u^2}{2} \Sigma_2,
\end{aligned}$$

then

$$\begin{aligned}
\Pr\{\lambda_d(\sum_k X_k) \leq \lambda_d(\sum_k \mathbb{E}X_k) - \epsilon\} &\leq \inf_{u \geq 0} \left\{ \exp(u\lambda_d(\sum_k \mathbb{E}X_k) - u\epsilon) \text{tr} \exp(\sum_k \log \mathbb{E} \exp(-uX_k)) \right\} \\
&\leq \inf_{u \geq 0} \left\{ d \exp(-u\epsilon + \frac{nu^2}{2} \lambda_1(\Sigma_2)) \right\} \\
&\leq d \exp(-\theta),
\end{aligned}$$

where $\epsilon = \sqrt{2\theta n \lambda_1(\Sigma_2)}$. □

Then we prove the Bernstein's moment condition for $\xi\xi^\top$ and $\xi\xi^\top - C$.

Lemma 4. *Let ξ be random vectors from $\mathcal{N}_d(0, C)$. For $p \geq 2$,*

$$\begin{aligned}
\mathbb{E}(\xi\xi^\top)^p &\preceq \frac{p!}{2} B^{p-2} (\text{tr}(C)C + 2C^2), \\
\mathbb{E}(\xi\xi^\top - C)^p &\preceq \frac{p!}{2} B^{p-2} \Sigma_2, \\
\mathbb{E}(C - \xi\xi^\top)^p &\preceq \frac{p!}{2} B^{p-2} \Sigma_2,
\end{aligned}$$

where $\Sigma_2 = \text{tr}(C)C + C^2$ and $B = 2\text{tr}(C)$.

Proof. Let $X = \xi\xi^\top$ and $\Sigma_p = \mathbb{E}(X - C)^p$, for $p \geq 2$. It follows Isserlis' theorem (Isserlis, 1918) that

$$\begin{aligned}
(\mathbb{E}X^2)_{ij} &= \sum_k \mathbb{E}\xi_i \xi_k^2 \xi_j = [\mathbb{E}\xi_i \xi_j] [\sum_k \mathbb{E}\xi_k^2] + 2 \sum_k [\mathbb{E}\xi_{ik}] [\mathbb{E}\xi_{jk}] = \text{tr}(C)C_{ij} + 2(C^2)_{ij}, \\
\Sigma_2 &= \mathbb{E}X^2 - C^2 = \text{tr}(C)C + C^2.
\end{aligned}$$

Then, we calculate $\mathbb{E}X^3$ and Σ_3 to get the basic idea.

$$\begin{aligned}
\mathbb{E}X^3 &= \text{tr}(C)^2C + 2\text{tr}(C^2)C + 4\text{tr}(C)C^2 + 8C^3 \preceq 5\text{tr}(C)(\text{tr}(C)C + 2C^2) \preceq \frac{3!}{2}B^{3-1}(\text{tr}(C)C + 2C^2), \\
\Sigma_3 &= \mathbb{E}X^3 - \mathbb{E}XCX - \mathbb{E}X^2C - \mathbb{E}CX^2 + \mathbb{E}C^2X + \mathbb{E}CXC + \mathbb{E}XC^2 - C^3 \\
&= \mathbb{E}X^3 - \mathbb{E}XCX - 2(\text{tr}(C)C^2 + C^3) \\
&= (\text{tr}(C)^2C + 2\text{tr}(C^2)C + 4\text{tr}(C)C^2 + 8C^3) - (\text{tr}(C^2)C + 2C^3) - 2(\text{tr}(C)C^2 + C^3) \\
&= \text{tr}(C)^2C + \text{tr}(C^2)C + 2\text{tr}(C)C^2 + 4C^3 \preceq 4\text{tr}(C)(\text{tr}(C)C + C^2) \preceq \frac{3!}{2}B^{3-1}\Sigma_2,
\end{aligned}$$

$$\mathbb{E}(C - X)^3 = -\Sigma_3 \preceq 0 \preceq \frac{3!}{2}B^{3-1}\Sigma_2.$$

Let $Z_{k,i} = \prod_j Y_{k,i,j}$, where $Y_{k,i,j}$ is X or C , k is the number of C 's in the term between 0 and p , i is the index term between 1 and $\binom{p}{k}$, and j is between 1 and p . Each element of $Y_{k,i,j}$ can be written as $\xi_{l_{j-1}}\xi_{l_j}$ or C_{l_{j-1},l_j} , where l_j is between 1 and d . It follows Isserlis' theorem that the expectation of each element $\mathbb{E}Z_{k,i}$ is the sum of the product of the expectations of $\xi_l\xi_{l'}$ all combinations. For example, in $p = 3$, we write $Z_{1,2} = XCX$, then

$$\begin{aligned}
\mathbb{E}Z_{1,2} &= (\mathbb{E} \sum_{l_1, \dots, l_2} \xi_{l_0}\xi_{l_1}C_{l_1,l_2}\xi_{l_2}\xi_{l_3} : l_0, l_3 \in \{1 \cdots d\}) \\
&= (\sum_{l_1, l_2} [\mathbb{E}(\xi_{l_0}\xi_{l_1})C_{l_1,l_2}\mathbb{E}(\xi_{l_2}\xi_{l_3}) + \mathbb{E}(\xi_{l_0}\xi_{l_2})C_{l_1,l_2}\mathbb{E}(\xi_{l_1}\xi_{l_3}) + \mathbb{E}(\xi_{l_0}\xi_{l_3})C_{l_1,l_2}\mathbb{E}(\xi_{l_1}\xi_{l_2})] : l_0, l_3 \in \{1 \cdots d\}) \\
&= (\sum_{l_1, l_2} [C_{l_0,l_1}C_{l_1,l_2}C_{l_2,l_3} + C_{l_0,l_2}C_{l_1,l_2}C_{l_1,l_3} + C_{l_0,l_3}C_{l_1,l_2}C_{l_1,l_2}] : l_0, l_3 \in \{1 \cdots d\}) \\
&= [(01)(12)(23)] + [(02)(12)(13)] + [(03)(12)(12)], \tag{9}
\end{aligned}$$

$$= [(0123)] + [(0213)] + [(03)(121)] \tag{10}$$

$$= C^3 + C^3 + \text{tr}(C^2)C \tag{11}$$

In Eq (9), each C is written a pair, and each product as a list. In Eq (10), pairs are combined into one chain and serveral loops. Then in Eq (11), each chain is C^c , where c is the lenth of the chain, and each loop is $\text{tr}(C^l)$, where l is the length of the loop. In general, $\mathbb{E}Z_{k,i}$ is the sum of terms like $C^c \prod_j \text{tr}(C^{l_j})$.

We have $C^c \preceq \text{tr}(C)^{c-2}C^2 \preceq \text{tr}(C)^{c-1}C$, and $\text{tr}(C^l) \leq \text{tr}(C)^l$, so we only count the terms with singleton chain, i.e. $c = 1$, and all terms to bound the expectations with $\text{tr}(C)^{p-2}(\text{tr}(C)C + 2C^2)$ or $\text{tr}(C)^{p-2}(\text{tr}(C)C + C^2)$. $\mathbb{E}Z_{k,i}$ is a expectation of $(2p - 2k)$ -order moments, which yields $(2p - 2k - 1)!!$ terms. For a given k , we have $\binom{p}{k}(2p - 2k - 1)!!$ terms, assuming $(-1)!! = 1$. A singleton chain term must contain $(0, p)$, thus $Z_{k,i}$ must be $X(\prod_{j=2}^{p-1} Y_{k,i,j})X$. For a given k , the number of singleton chain terms is $\binom{p-2}{k}(2p - 2k - 3)!!$.

For $\mathbb{E}X^p = \mathbb{E}Z_{0,1}$ has $(2p - 1)!!$ terms, which include $(2p - 3)!!$ singleton chain terms. The number of singleton chain terms is less than a third of the number of all terms when $p \geq 2$. For $p \geq 2$,

$$\begin{aligned}
\mathbb{E}X^p &\preceq \frac{(2p - 1)!!}{3}(\text{tr}(C)^{p-1}C + 2\text{tr}(C)^{p-2}C^2) = \frac{\Gamma(p + 1/2)}{3\sqrt{\pi}\Gamma(p + 1)}p!2^p(\text{tr}(C)^{p-1}C + 2\text{tr}(C)^{p-2}C^2) \\
&\preceq \frac{1}{8}p!2^p(\text{tr}(C)^{p-1}C + 2\text{tr}(C)^{p-2}C^2) = \frac{p!}{2}B^{p-2}(\text{tr}(C)C + 2C^2).
\end{aligned}$$

Then $\mathbb{E}(X - C)^p = \sum_k (-1)^k \sum_i Z_{k,i}$. The number of singleton chain terms is less than half of the number of all terms. Thus

$$\begin{aligned}
\Sigma_4 &\preceq 10\text{tr}(C)^{4-1}C + 50\text{tr}(C)^{4-2}C^2 \preceq 30\text{tr}(C)^{4-1}C + 30\text{tr}(C)^{4-2}C^2 \preceq \frac{4!}{2}B^{4-2}\Sigma_2, \\
\mathbb{E}(C - X)^4 &= \Sigma_4 \preceq \frac{4!}{2}B^{4-2}\Sigma_2.
\end{aligned}$$

When $p \geq 5$,

$$\begin{aligned}
\Sigma_p &\preceq \mathbb{E}X^p + C^p \preceq \frac{(2p-1)!! + 1}{2}(\text{tr}(C)^{p-1}C + \text{tr}(C)^{p-2}C^2) \\
&= \left(\frac{\Gamma(p+1/2)}{2\sqrt{\pi}\Gamma(p+1)} + \frac{1}{p!2^{p+1}}\right)p!2^p\text{tr}(C)^{p-2}(\text{tr}(C)C + C^2) \\
&\preceq 0.1232 \times p!2^p\text{tr}(C)^{p-2}(\text{tr}(C)C + C^2) \preceq \frac{p!}{2}B^{p-2}\Sigma_2, \\
\mathbb{E}(C - X)^p &\preceq \mathbb{E}X^p + C^p \preceq \frac{p!}{2}B^{p-2}\Sigma_2.
\end{aligned}$$

□

Now we prove Theorem 1.

Proof of Theorem 1. Let $X_k = \xi_k \xi_k^\top - C$. We have $\mathbb{E}X_k = 0$, $\lambda_1(\Sigma_2) \leq (r+1)\lambda_1(C)^2$, and $B = 2r\lambda_1(C)$. Then Eq (1) follows Lemma 4 and Theorem 3. Similarly, letting $X_k = C - \xi_k \xi_k^\top$, we prove Eq (2). Combining them and $\|C\| = \lambda_1(C)$, we have Eq (3). Plugging $\lambda_1(C) = \kappa_\ell \lambda_\ell(C)$, Eq (4) follows Weyl's theorem on eigenvalues, specifically,

$$\begin{aligned}
\lambda_\ell\left(\frac{1}{n}S\right) &\leq \lambda_\ell(C) + \lambda_1\left(\frac{1}{n}S - C\right), \\
\lambda_\ell(C) &\leq \lambda_\ell\left(\frac{1}{n}S\right) + \lambda_1\left(C - \frac{1}{n}S\right).
\end{aligned}$$

□

References

- Birgé, L., & Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4, 329–375.
- Gittens, A., & Tropp, J. A. (2011). Tail bounds for all eigenvalues of a sum of random matrices. arXiv:1104.4513.
- Isserlis, L. (1918). On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12, 134–139.
- Laurent, B., & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28, 1302–1338.
- Tropp, J. A. (2010). User-friendly tail bounds for sums of random matrices. arXiv:1004.4389.